Spatial imputation for air pollutants data sets via low rank matrix completion algorithm

Environment International 139 (2020) 105713

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Spatial imputation for air pollutants data sets via low rank matrix completion algorithm

Xiaofeng Liu[a,b,c,*], Xue Wang[a,b], Lang Zou[a,f], Jing Xia[d], Wei Pang[e]

[a] College of IoT Engineering, Hohai University, Changzhou 213022, China
[b] School of Information and Engineering, Changzhou University, Changzhou 213164, China
[c] Jiangsu Key laboratory of Special Robot Technology, Changzhou 213022, China
[d] Changzhou Environmental Monitoring Center, Changzhou 213022, China
[e] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK
[f] Huawei Nanjing Research Institute, China

ARTICLE INFO

Handling Editor: Xavier Querol

Keywords:
Air pollutants
Low rank matrix completion
Missing data
Spatial imputation

ABSTRACT

Incomplete observation of hourly air-pollutants concentration data is a common issue existing in urban air quality monitoring networks. This research proposes a spatial interpolation method to impute missing values presented in air pollutants data sets based on low rank matrix completion (LRMC). It considers air pollutants data of high correlation and consistency in its spatial distribution. We evaluate the performance of the proposed method when imputing various air pollutants concentration time series (NO_x, O_3, SO_2, $PM_{2.5}$, PM_{10}) in terms of root mean square error (*RMSE*), index of agreement (D^2), and goodness of fit (R^2). It systematically compared with existing established imputation techniques, including nearest neighboring, mean substitution, regression-based method, spline interpolation, spectral method, and regularized expectation maximization algorithm (EM). As a spatial imputation method, LRMC outperforms these methods used in this research under the condition of larger missing ratio (such as 30% removal) on the central air pollutants monitoring station. For all monitoring stations, comprehensive experimental results show that LRMC always generates robust results to replace missing data with reasonable substitutions, and it is not sensitive to the length of missing gaps. The promising imputation performance in terms of the indicator R^2 obtained by the proposed LRMC demonstrates that it can effectively impute missing values of air pollutants time series based on their inherent patterns.

1. Introduction

Air pollution poses potential risks to human health (Listed, 2014). Air pollutants such as airborne particulate matter and ozone have been associated with increases in mortality and hospital admissions due to their correlations with respiratory and cardiovascular diseases (Brunekreef and Holgate, 2002). Air quality monitoring (AQM) network is a major facility for assessing outdoor air pollution conditions and developing air pollution control plans (Baldauf et al., 2001). Despite the quality assurance and quality control procedures (von Lehmden et al., 1979; H and Beebe, 1985), hourly air pollutants concentration data received from AQM stations are often presented with some missing values, which brings great impediment to data-enabled applications such as online air quality publishing, ensemble forecasting and epidemiological studies (Wang et al., 2014; Chandrappa and Kulshrestha, 2016; Pope, 2000). Failing to acquire continuous data can be caused by machine errors, regular maintenance, or power cuts. Imputation for air pollutants time series is an essential task, especially when the missing ratio is beyond limit of tolerance (Mansour et al., 2014).

Techniques available for imputing missing data can be divided into two main categories: single imputation and multiple imputation. Single imputation methods replace each missing value with a precise value. The complete data then can be directly applied to interpret results in related research fields. Multiple imputation methods generate multiple simulated values for each missing one, in order to reflect the uncertainty attached to the missing data (Shafer, 1997). Generally a multiple imputation method requires a full assumption of the distributional form of variable in order to derive the conditional distribution of the missing data given the observed data. Air pollutants data sets are always in the form of matrix, where each column (variable) can be time series of various air pollutants (NO_x, O_3, $PM_{2.5}$, SO_2 etc.), meteorological factors (wind speed, temperature, humidity and pressure), or data from other monitoring stations. Imputation of missing air pollutants data is a challenging task because the creation and

* Corresponding author att: College of IoT Engineering, Hohai University, Changzhou 213022, China.
E-mail addresses: xfliu@hhu.edu.cn (X. Liu), wangxue@cczu.edu.cn (X. Wang).

https://doi.org/10.1016/j.envint.2020.105713
Received 9 November 2019; Received in revised form 2 March 2020; Accepted 31 March 2020

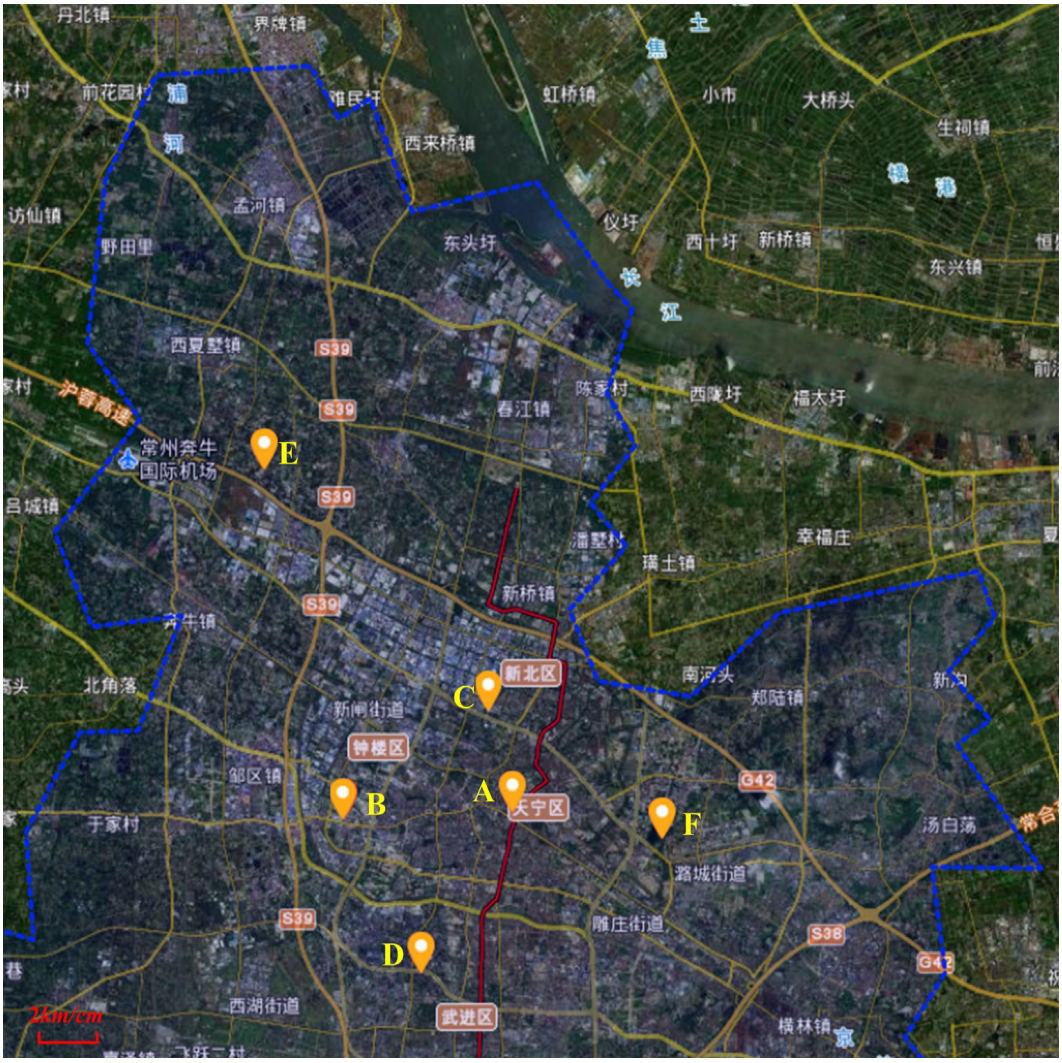

Fig. 1. Spatial distribution of air quality monitoring stations in Changzhou, China. The map scale is 2 km/cm. Multiple stations can provide high spatial resolution of air pollution. (A: Central, B: ZhouLou, C: XinBei, D: WuJin, E: AnJia, F: JinKai).

Table 1
Detailed information of air-pollutants experimental dataset with missing data.

Station	O_3	SO_2	PM_{10}	$PM_{2.5}$	NO_x
	number-of-gaps (max-gap)				
Central(A)	31(16)	27(5)	47(10)	16(15)	28(7)
ZhongLou(B)	28(20)	21(18)	14(18)	10(18)	13(8)
XinBei(C)	16(6)	23(6)	12(9)	3(6)	21(8)
WuJin(D)	30(5)	15(2)	11(2)	11(13)	15(8)
AnJia(E)	27(11)	15(4)	23(10)	12(31)	12(5)
JinKai(F)	29(34)	30(8)	18(10)	11(8)	41(26)
Station	univariate statistical mean(mg/m^3)				
Central(A)	0.062	0.019	0.100	0.060	0.040
ZhongLou(B)	0.078	0.011	0.101	0.051	0.052
XinBei(C)	0.096	0.017	0.073	0.082	0.044
WuJin(D)	0.070	0.025	0.126	0.072	0.046
AnJia(E)	0.084	0.010	0.184	0.054	0.036
JinKai(F)	0.066	0.013	0.164	0.077	0.053
missing-ratio	1.13%	0.85%	0.92%	0.06%	0.22%

Table 2
Spatial correlation analysis of air pollutants data from multiple monitoring stations.

S-S(d/km)	global correlation coefficient(ρ)				
	O_3	SO_2	PM_{10}	$PM_{2.5}$	NO_x
A ↔ C(4.6)	0.9564	0.7408	0.9261	0.9501	0.8555
A ↔ F(6.8)	0.9303	0.7323	0.9046	0.9257	0.7916
A ↔ B(7.4)	0.9485	0.8637	0.9460	0.9628	0.8887
B ↔ D(7.4)	0.9390	0.8019	0.9264	0.9401	0.7486
B ↔ C(7.9)	0.9562	0.7043	0.9466	0.9619	0.8576
A ↔ D(8.0)	0.9164	0.8258	0.9047	0.9342	0.7563
C ↔ F(8.8)	0.9390	0.6973	0.9202	0.9288	0.8039
D ↔ F(11.9)	0.8949	0.6994	0.8861	0.8969	0.7484
C ↔ D(12.9)	0.9180	0.6714	0.9176	0.9234	0.7167
C ↔ E(14.0)	0.9297	0.6588	0.9300	0.9359	0.7914
B ↔ E(14.4)	0.9127	0.7304	0.9155	0.9250	0.7717
B ↔ F(14.6)	0.9173	0.7296	0.9013	0.9121	0.8132
A ↔ E(17.4)	0.8991	0.7025	0.8959	0.9103	0.7683
E ↔ F(22.5)	0.9075	0.6729	0.8993	0.9066	0.7427
D ↔ E(24.3)	0.8695	0.6733	0.8882	0.8914	0.6253

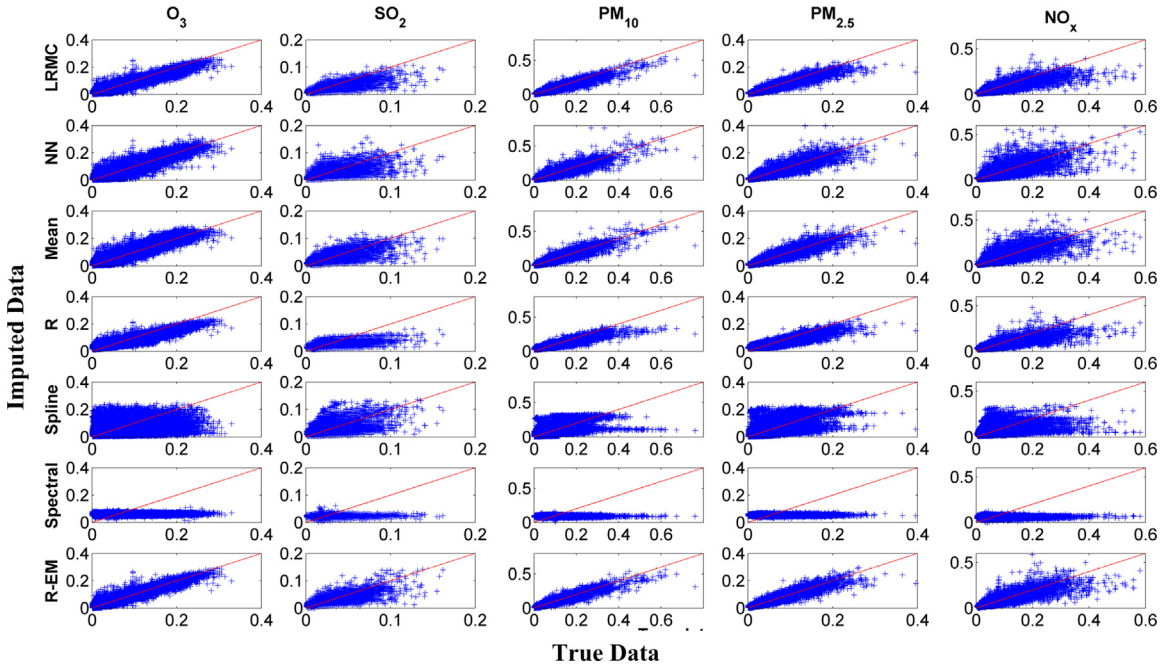

Fig. 2. Spatial imputation results of various air pollutants using LRMC algorithm in form of scatter plot of value pair between true values and corresponding imputed values. In each figure, $52704 = 24 \times 366 \times 6$ points are involved in the experimental case of missing gap length $l = 100$. The red straight line is an angular bisectors between x-axis and y-axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Imputation performance under different missing mechanisms and missing ratios.

Indicators	Pollutants	NN	Mean	Regression	Spine	Spectral	R-EM	LRMC
RMSE								
5%	NO_x	0.029	0.028	0.029	0.027	0.056	**0.025**	**0.025**
	O_3	0.018	0.014	0.018	0.016	0.049	**0.009**	0.011
	SO_2	0.012	0.006	0.007	0.006	0.013	**0.005**	**0.005**
	PM_{10}	0.022	0.018	0.021	0.022	0.058	**0.016**	0.017
	$PM_{2.5}$	0.012	0.010	0.014	0.012	0.038	**0.009**	0.011
10%	NO_x	0.038	0.035	0.028	0.027	0.059	0.026	**0.025**
	O_3	0.018	0.015	0.018	0.016	0.047	**0.014**	**0.014**
	SO_2	0.013	0.012	0.010	0.006	0.018	0.009	**0.006**
	PM_{10}	0.028	0.024	0.023	0.021	0.063	0.019	**0.018**
	$PM_{2.5}$	0.016	0.015	0.016	0.014	0.040	0.012	**0.011**
D^2								
5%	NO_x	0.910	0.910	0.913	0.933	0.505	**0.944**	0.931
	O_3	0.965	0.977	0.957	0.969	0.557	**0.989**	0.985
	SO_2	0.819	0.927	0.895	0.939	0.556	**0.954**	0.951
	PM_{10}	0.961	0.971	0.958	0.960	0.592	**0.978**	0.975
	$PM_{2.5}$	0.973	0.981	0.957	0.970	0.560	**0.984**	**0.984**
10%	NO_x	0.875	0.873	0.928	0.933	0.422	**0.945**	0.931
	O_3	0.961	0.970	0.952	0.969	0.480	0.963	**0.973**
	SO_2	0.803	0.826	0.903	0.939	0.411	0.926	**0.935**
	PM_{10}	0.942	0.956	0.958	0.960	0.463	**0.964**	0.959
	$PM_{2.5}$	0.957	0.961	0.949	0.970	0.456	0.964	**0.972**
R^2								
5%	NO_x	0.723	0.796	0.729	0.767	0.420	0.802	**0.828**
	O_3	0.908	0.922	0.876	0.891	0.402	**0.960**	0.955
	SO_2	0.535	0.755	0.677	0.785	0.311	0.833	**0.850**
	PM_{10}	0.855	0.897	0.858	0.852	0.326	0.917	**0.925**
	$PM_{2.5}$	0.900	0.932	0.856	0.889	0.381	0.941	**0.949**
10%	NO_x	0.603	0.711	0.766	0.767	0.344	0.802	**0.809**
	O_3	0.858	0.893	0.865	0.891	0.370	0.938	**0.942**
	SO_2	0.537	0.634	0.695	0.785	0.343	0.744	**0.842**
	PM_{10}	0.818	0.874	0.858	0.852	0.301	0.903	**0.914**
	$PM_{2.5}$	0.856	0.884	0.831	0.889	0.322	0.903	**0.926**

Table 4
Imputation performance under different missing mechanisms and missing ratios for the central monitoring station A.

Indicators	Pollutants	NN	Mean	Regression	Spine	Spectral	EM	LRMC
RMSE	O_3	0.0195	0.0135	0.0291	0.0304	0.0463	0.0126	**0.0105**
5%	SO_2	0.0116	0.0075	0.0078	0.0079	0.0113	0.0064	**0.0047**
	PM_{10}	0.0297	0.0251	0.0284	0.0599	0.0760	0.0196	**0.0186**
	$PM_{2.5}$	0.0118	0.0093	0.0191	0.0214	0.0369	0.0092	**0.0089**
	No_x	0.0223	0.0217	0.0240	0.0546	0.0488	**0.0206**	0.0211
10%	O_3	0.0206	0.0147	0.0275	0.0324	0.0492	0.0123	**0.0118**
	SO_2	0.0107	0.0068	0.0076	0.0091	0.0110	0.0057	**0.0047**
	PM_{10}	0.0249	0.0211	0.0261	0.0516	0.0647	0.0172	**0.0169**
	$PM_{2.5}$	0.0106	0.0086	0.0182	0.0201	0.0354	0.0084	**0.0079**
	No_x	0.0235	0.0233	0.0258	0.0517	0.0506	**0.0202**	0.0211
20%	O_3	0.0203	0.0149	0.0284	0.0363	0.0496	0.0132	**0.0119**
	SO_2	0.0107	0.0068	0.0085	0.0112	0.0128	0.0059	**0.0052**
	PM_{10}	0.0236	0.0202	0.0262	0.0497	0.0651	0.0179	**0.0176**
	$PM_{2.5}$	0.0117	0.0096	0.0190	0.0237	0.0400	0.0094	**0.0091**
	No_x	0.0255	0.0239	0.0289	0.0512	0.0554	**0.0207**	0.0217
25%	O_3	0.0198	0.0145	0.0298	0.0313	0.0478	0.0137	**0.0120**
	SO_2	0.0107	0.0067	0.0080	0.0125	0.0125	0.0055	**0.0049**
	PM_{10}	0.0240	0.0204	0.0246	0.0513	0.0631	**0.0174**	0.0179
	$PM_{2.5}$	0.0121	0.0098	0.0175	0.0244	0.0383	0.0095	**0.0091**
	No_x	0.0269	0.0265	0.0272	0.0451	0.0551	0.0241	**0.0228**
33%	O_3	0.0199	0.0152	0.0267	0.0337	0.0501	**0.0119**	0.0120
	SO_2	0.0129	0.0067	0.0083	0.0117	0.0128	0.0052	**0.0051**
	PM_{10}	0.0223	0.0179	0.0229	0.0485	0.0600	0.0156	**0.0154**
	$PM_{2.5}$	0.0116	0.0094	0.0179	0.0238	0.0388	0.0093	**0.0090**
	No_x	0.0269	0.0261	0.0310	0.0492	0.0551	0.0272	**0.0240**
D2	O_3	0.9557	0.9758	0.8639	0.8821	0.2060	0.9788	**0.9853**
5%	SO_2	0.7896	0.8838	0.8188	0.8823	0.2660	0.9125	**0.9479**
	PM_{10}	0.9647	0.9725	0.9567	0.8954	0.2102	0.9818	**0.9828**
	$PM_{2.5}$	0.9732	0.9825	0.9066	0.9263	0.2092	0.9828	**0.9830**
	No_x	0.9199	0.9201	0.8940	0.6431	0.2877	**0.9434**	0.9219
10%	O_3	0.9579	0.9749	0.8953	0.8866	0.2059	0.9828	**0.9834**
	SO_2	0.7950	0.8938	0.8291	0.8204	0.2737	0.9225	**0.9416**
	PM_{10}	0.9630	0.9717	0.9474	0.8822	0.2193	**0.9801**	**0.9801**
	$PM_{2.5}$	0.9767	0.9839	0.9108	0.9234	0.2097	0.9850	**0.9861**
	No_x	0.9130	0.9106	0.8919	0.7219	0.2787	**0.9488**	0.9282
20%	O_3	0.9604	0.9745	0.8856	0.8609	0.1693	0.9806	**0.9835**
	SO_2	0.8468	0.9235	0.8511	0.8179	0.1971	0.9425	**0.9488**
	PM_{10}	0.9652	0.9734	0.9466	0.8680	0.2320	**0.9785**	0.9779
	$PM_{2.5}$	0.9776	0.9844	0.9266	0.9116	0.1712	0.9850	**0.9853**
	No_x	0.9229	0.9292	0.8949	0.7727	0.2555	**0.9558**	0.9420
25%	O_3	0.9586	0.9742	0.8613	0.8921	0.1685	0.9767	**0.9820**
	SO_2	0.8363	0.9185	0.8600	0.8061	0.1817	0.9467	**0.9518**
	PM_{10}	0.9603	0.9707	0.9502	0.8558	0.2499	**0.9778**	0.9753
	$PM_{2.5}$	0.9735	0.9820	0.9336	0.9050	0.1792	0.9833	**0.9839**
	No_x	0.9096	0.9086	0.9064	0.8043	0.2647	**0.9396**	0.9351
33%	O_3	0.9622	0.9740	0.9010	0.8926	0.1693	**0.9847**	0.9835
	SO_2	0.8070	0.9274	0.8564	0.7991	0.1871	**0.9549**	0.9514
	PM_{10}	0.9623	0.9752	0.9538	0.8441	0.2494	**0.9809**	0.9803
	$PM_{2.5}$	0.9764	0.9841	0.9338	0.9061	0.1715	0.9847	**0.9848**
	No_x	0.9154	0.9154	0.8788	0.7914	0.2468	0.9236	**0.9302**
R2	O_3	0.8833	0.9178	0.6123	0.6154	0.0008	0.9292	**0.9541**
5%	SO_2	0.5051	0.6483	0.4936	0.6299	0.0016	0.7051	**0.8212**
	PM_{10}	0.8901	0.9013	0.8551	0.8009	0.0001	0.9300	**0.9494**
	$PM_{2.5}$	0.8981	0.9353	0.7357	0.7756	0.0001	0.9367	**0.9541**
	No_x	0.7449	0.8113	0.7142	0.1987	0.0001	0.7988	**0.8280**
10%	O_3	0.8905	0.9135	0.6907	0.6284	0.0011	0.9373	**0.9505**
	SO_2	0.4721	0.6617	0.5043	0.4719	0.0013	0.7309	**0.8065**
	PM_{10}	0.8723	0.8962	0.8255	0.7267	0.0000	0.9237	**0.9380**
	$PM_{2.5}$	0.9110	0.9397	0.7298	0.7486	0.0002	0.9421	**0.9561**
	No_x	0.7452	0.8150	0.6866	0.3134	0.0000	0.8143	**0.8376**
20%	O_3	0.8950	0.9130	0.6764	0.5594	0.0005	0.9307	**0.9507**
	SO_2	0.5722	0.7431	0.5558	0.4777	0.0010	0.7930	**0.8441**

(*continued on next page*)

Table 4 (*continued*)

Indicators	Pollutants	NN	Mean	Regression	Spine	Spectral	EM	LRMC
	PM_{10}	0.8733	0.9024	0.8273	0.6201	0.0000	0.9183	**0.9334**
	$PM_{2.5}$	0.9148	0.9417	0.7711	0.6999	0.0002	0.9431	**0.9551**
	No_x	0.7619	0.8517	0.6876	0.3920	0.0000	0.8395	**0.8702**
25%	O_3	0.8903	0.9120	0.6132	0.6415	0.0005	0.9203	**0.9441**
	SO_2	0.5444	0.7288	0.5790	0.4971	0.0004	0.8075	**0.8527**
	PM_{10}	0.8533	0.8913	0.8392	0.6005	0.0001	0.9175	**0.9301**
	$PM_{2.5}$	0.9012	0.9322	0.7904	0.6952	0.0000	0.9366	**0.9535**
	No_x	0.7358	0.8072	0.7212	0.4355	0.0000	0.7853	**0.8490**
33%	O_3	0.8965	0.9116	0.7231	0.6483	0.0006	0.9445	**0.9523**
	SO_2	0.5292	0.7569	0.5728	0.4360	0.0003	0.8345	**0.8473**
	PM_{10}	0.8610	0.9098	0.8407	0.5564	0.0000	0.9266	**0.9403**
	$PM_{2.5}$	0.9102	0.9400	0.7820	0.6884	0.0001	0.9410	**0.9527**
	No_x	0.7278	0.7831	0.6336	0.4355	0.0000	0.7349	**0.8065**

dispersion of air pollutants exhibit highly variation in spatial and temporal patterns due to uncertain pollution sources and the complexity of the physicochemical processes (Seinfeld and Pandis, 2006).

The univariate imputation methods for air pollutants data use the temporal structure such as periodicity, trend and autocorrelation (Box, 2008). For instance, temporal substitution uses historical data of the same periods to replace the missing or replaces the missing air pollutants data with the mean of neighboring values (Noor, 2015; Plaia and Bondì, 2006). Temporal substitution is relatively simple but does not consider any temporal variation of air pollutants data. It is generally used in applications with low requirements for data quality (Junninen et al., 2004). Temporal interpolation (e.g., linear or cubic interpolation) uses a straight line or curve to fit the observed data, from which the unobserved values can be estimated (Shukri et al., 2008). The filled values obtained by interpolation methods can well simulate the temporal trend of air pollutants data. However, the bias between imputed values and the corresponding true values increases significantly when data are missing in longer gaps since the data structure during an unobserved period is unpredictable from an univariate view. Spectral methods for univariate imputation consider that the observation sequence of air pollutant variables is band-limited. It can be implemented via time–frequency transform techniques (e.g., discrete cosine transform or discrete Fourier transform) in an iterative procedure by zeroing high frequency coefficients (Moshenberg et al., 2015). The missing values can be reconstructed with minimum error under the discrete sampling theorem(Yaroslavsky et al., 2009). Spectral methods can well recover values in the unobserved periods where the temporal variation of air pollutant concentration is relatively slow. The drawback of spectral method is obvious that they use global computation of the complete data, which underestimates local changeability and non-stationarity of air pollutants time series.

Another technique for data imputation is multivariate imputation, typical regression-based methods and expectation maximization (EM) algorithm, which rely on correlations between different variables in order to estimate values for the missing data (Shahbazi et al., 2018; Junger and Leon, 2015). Regression-based methods predict the missing values using other variables as predictors. The common limitation of regression model is that the parameters in model are sensitive to sampling and irregular values (outliers or anomalies presented in air pollutants datasets) that are hard to be well-eliminated via the model itself. EM algorithm is an iterative procedure that produces maximum likelihood estimate for missing data, which has two key steps: for the E-step at one iteration, if the value is missing, the best substitution is calculated from a posterior expectation of missing values based on current parameters (means, variances and covariances). In the M-step of the same iteration, the parameters are re-estimated using the current fully observed values based on maximum likelihood principle. EM algorithm contains much computation of matrix inversion in the iterative steps. The result can be incorrect when the covariance matrix is ill-conditioned, which can be solved by regularization method (Schneider, 2001). More methods applied to air pollutants data imputation are summarised and discussed in (Gómez-Carracedo et al., 2014).

Although the aforementioned methods often produce good estimates for missing values, the inherent data structure such as correlation, trends and seasonality are always underestimated or overestimated (Box, 2008). This research aims to provide an imputation method that replace missing value with reasonable substitutions. A spatial imputation scheme, comprehensive utilization of data from multiple monitoring stations, is presented using low rank matrix completion algorithm. It is inspired by high correlation and consistency of the spatial distribution of air pollutants data. Spatial correlation means data from geographically adjacent sites have similar temporal trends. Spatial consistency implies synchronous measurements (e.g. $PM_{2.5}$, PM_{10}) should not vary greatly although outliers sometimes occur in air quality data sets (Guan, 2016). In the following section, we will elaborate the procedure of LRMC-based spatial imputation and give some discussions on the performance to address the missing air pollutants data.

2. Methods

2.1. Spatial imputation based Low rank matrix completion algorithm

Low rank matrix completion algorithm (Shang et al., 2014; Candès and Recht, 2009) reconstructs a matrix from the observed subset of its entries based on the low rank property of the original matrix. It fills the unobserved data that makes the matrix has lower rank with the observed data unchanged. Given an incomplete spatial matrix of air pollutants data D with some missing values, we define the missing indicator matrix M (the same size of D), of which each element is 0 if the corresponding position of D is missing and 1 if observed. The low-rank matrix completion algorithm can be described by:

$$\begin{aligned} &\min \quad rank(A) \\ &s.t. \quad M^{\circ}A = M^{\circ}D, \end{aligned} \tag{1}$$

where A is the recovered low-rank matrix. The symbol $^{\circ}$ is the element-wise product of two matrices. To address this minimum optimization problem, the rank of matrix A is a non-derivable term and we always use the nuclear norm as its convex approximation. Therefore, the problem is converted to minimize $\|A\|_*$ under the same constraint $M^{\circ}A = M^{\circ}D$, which can be solved using Lagrange multiplier method. In this research, we consider air pollutants data being presented by additive sparse noise resulting from occasional errors of sensors and outliers caused by extreme weather. Thus, matrix D is decomposed to the sum of a low-rank matrix A and a sparse matrix E (Wright et al., 2009). The low rank based imputation can be formulated as:

Table 5
Mean and standard deviation values of imputation indicators of boundary station B~F.

Indicators	Pollutants	NN	Mean	Regression	Spine	EM	LRMC
RMSE	O_3	0.0195±0.0017	0.0177±0.0038	0.0300±0.0028	0.0305±0.0057	0.0154±0.0023	**0.0149±0.0021**
5%	SO_2	0.0111±0.0049	0.0108±0.0037	0.0105±0.0044	0.0152±0.0064	0.0094±0.0047	**0.0090±0.0048**
	PM_{10}	0.0252±0.0052	**0.0180±0.0026**	0.0271±0.0034	0.0472±0.0063	0.0189±0.0040	0.0208±0.0019
	$PM_{2.5}$	0.0127±0.0009	0.0104±0.0021	0.0183±0.0009	0.0247±0.0053	0.0101±0.0027	**0.0100±0.0020**
	No_x	0.0274±0.0036	0.0249±0.0063	0.0225±0.0056	0.0371±0.0096	0.0206±0.0065	**0.0189±0.0059**
10%	O_3	0.0198±0.0009	0.0176±0.0027	0.0288±0.0021	0.0314±0.0055	**0.0152±0.0019**	0.0153±0.0021
	SO_2	0.0101±0.0037	0.0097±0.0025	0.0097±0.0031	0.0138±0.0050	0.0084±0.0032	**0.0080±0.0034**
	PM_{10}	0.0232±0.0032	**0.0176±0.0031**	0.0254±0.0017	0.0451±0.0053	0.0181±0.0030	0.0193±0.0019
	$PM_{2.5}$	0.0123±0.0014	0.0106±0.0025	0.0180±0.0014	0.025±0.00650	**0.0099±0.0025**	0.0101±0.0024
	No_x	0.0302±0.0038	0.0249±0.0065	0.0226±0.0050	0.0437±0.0166	0.0198±0.0051	**0.0192±0.0053**
20%	O_3	0.0190±0.0007	0.0165±0.0025	0.0284±0.0018	0.0312±0.0036	0.0145±0.0019	**0.0142±0.0017**
	SO_2	0.0106±0.0030	0.0100±0.0019	0.0097±0.0024	0.0145±0.0048	0.0083±0.0025	**0.0080±0.0027**
	PM_{10}	0.0238±0.0036	0.0190±0.0043	0.0252±0.0023	0.0437±0.0047	**0.0184±0.0030**	0.0193±0.0030
	$PM_{2.5}$	0.0132±0.0016	0.0116±0.0027	0.0186±0.0019	0.0242±0.0047	0.0108±0.0026	**0.0107±0.0024**
	No_x	0.0293±0.0022	0.0264±0.0054	0.0236±0.0047	0.0447±0.0125	**0.0203±0.0053**	0.0205±0.0061
25%	O_3	0.0192±0.0003	0.0164±0.0022	0.0335±0.0023	0.0317±0.0035	0.0160±0.0024	**0.0141±0.0018**
	SO_2	0.0104±0.0026	0.0096±0.0018	0.0090±0.0020	0.0145±0.0038	0.0083±0.0026	**0.0077±0.0026**
	PM_{10}	0.0240±0.0029	0.0189±0.0035	0.0271±0.0019	0.0433±0.0036	**0.0176±0.0031**	0.0191±0.0031
	$PM_{2.5}$	0.0134±0.0017	0.0113±0.0024	0.0179±0.0018	0.0238±0.0028	0.0106±0.0026	**0.0106±0.0025**
	No_x	0.0290±0.0036	0.0257±0.0047	0.0286±0.0055	0.043±0.0128	0.0219±0.0068	**0.0206±0.0062**
D_2	O_3	0.9487±0.0158	0.9545±0.0217	0.8402±0.0199	0.8732±0.0379	0.9647±0.0112	**0.9663±0.0118**
5%	SO_2	0.7806±0.1285	0.8079±0.0972	0.7966±0.1056	0.7561±0.1396	**0.8584±0.1165**	0.8522±0.1198
	PM_{10}	0.9730±0.0125	**0.9865±0.0039**	0.9646±0.0107	0.9338±0.0186	0.9840±0.0077	0.9803±0.0046
	$PM_{2.5}$	0.9666±0.0056	0.9762±0.0097	0.9082±0.0108	0.8957±0.0389	0.9762±0.0125	**0.9768±0.0090**
	No_x	0.8397±0.0650	0.8698±0.0745	0.8818±0.0304	0.7787±0.0671	**0.9135±0.0312**	0.9193±0.0293
10%	O_3	0.9527±0.0086	0.9615±0.0127	0.8676±0.0185	0.8815±0.0283	**0.9691±0.0103**	0.9686±0.0104
	SO_2	0.7882±0.0934	0.8186±0.0613	0.8054±0.0515	0.7530±0.0892	**0.8744±0.0632**	0.8680±0.0722
	PM_{10}	0.9668±0.0100	**0.9807±0.0066**	0.9542±0.0083	0.9083±0.0200	0.9787±0.0085	0.9750±0.0060
	$PM_{2.5}$	0.9664±0.0067	0.9738±0.0115	0.9093±0.0079	0.8850±0.0401	**0.9767±0.0111**	0.9751±0.0107
	No_x	0.7972±0.1068	0.8624±0.0880	0.8780±0.02890	0.7433±0.1013	**0.9190±0.0279**	0.9121±0.0287
20%	O_3	0.9577±0.0045	0.9665±0.0106	0.8676±0.0158	0.8868±0.0109	0.9724±0.0096	**0.9728±0.0093**
	SO_2	0.8195±0.0631	0.8503±0.0339	0.8366±0.0288	0.7863±0.0538	**0.9013±0.0393**	0.8970±0.0441
	PM_{10}	0.9634±0.0099	0.9760±0.0100	0.9527±0.0080	0.9006±0.0190	**0.9772±0.0079**	0.9740±0.0076
	$PM_{2.5}$	0.9694±0.0061	0.9754±0.0104	0.9258±0.0094	0.9057±0.0270	**0.9787±0.0087**	0.9781±0.0087
	No_x	0.8624±0.0525	0.8944±0.0408	0.9046±0.0085	0.7809±0.0528	**0.9395±0.0162**	0.9285±0.0226
25%	O_3	0.9537±0.0036	0.9647±0.0094	0.8338±0.0263	0.8794±0.0133	0.9631±0.0127	**0.9718±0.009**
	SO_2	0.8183±0.0512	0.8522±0.0313	0.8439±0.0230	0.7672±0.0290	0.8979±0.0319	**0.8995±0.0376**
	PM_{10}	0.9596±0.0082	0.9745±0.0087	0.9544±0.0086	0.8919±0.0095	**0.9775±0.0077**	0.9722±0.0079
	$PM_{2.5}$	0.9657±0.0077	0.9747±0.0101	0.9308±0.0126	0.9050±0.0168	**0.9776±0.0097**	0.9767±0.0098
	No_x	0.8642±0.0275	0.8924±0.0370	0.8990±0.0192	0.7866±0.0512	**0.925±0.0245**	0.9217±0.0320
R_2	O_3	0.8411±0.0647	0.8627±0.0718	0.5557±0.0441	0.5956±0.1051	0.8778±0.0424	**0.8991±0.0419**
5%	SO_2	0.5244±0.1758	0.5827±0.1737	0.5131±0.1293	0.4548±0.2145	0.6383±0.1898	**0.6615±0.1937**
	PM_{10}	0.9067±0.0401	0.9524±0.0179	0.9036±0.0263	0.842±0.0496	0.9460±0.0260	**0.9581±0.0136**
	$PM_{2.5}$	0.8858±0.0176	0.9169±0.0352	0.7427±0.0439	0.7026±0.0909	0.9129±0.0455	**0.9326±0.0341**
	No_x	0.5575±0.1438	0.6764±0.1462	0.6626±0.0793	0.4208±0.1368	0.7260±0.1004	**0.8014±0.0507**
10%	O_3	0.8563±0.0391	0.8872±0.0460	0.6256±0.0540	0.6153±0.0797	0.8911±0.0377	**0.9087±0.0341**
	SO_2	0.5141±0.1283	0.5885±0.1259	0.4763±0.0704	0.3919±0.1540	0.6336±0.1255	**0.6637±0.1281**
	PM_{10}	0.8836±0.0290	0.9317±0.0244	0.8631±0.0192	0.7487±0.0574	0.9242±0.0294	**0.9388±0.0196**
	$PM_{2.5}$	0.8841±0.0236	0.9099±0.0388	0.7327±0.0322	0.6611±0.0751	0.9143±0.0418	**0.9275±0.0353**
	No_x	0.4963±0.1891	0.6616±0.1667	0.6564±0.0895	0.4012±0.1336	0.7461±0.0981	**0.7773±0.0719**
20%	O_3	0.8663±0.0278	0.8992±0.0413	0.6364±0.0452	0.6283±0.0315	0.9015±0.0358	**0.9207±0.0302**
	SO_2	0.5667±0.0869	0.6437±0.1044	0.5406±0.0569	0.4407±0.1045	0.6857±0.0907	**0.7208±0.0986**
	PM_{10}	0.8720±0.0229	0.9173±0.0305	0.8533±0.0248	0.6982±0.0496	0.9161±0.0291	**0.9299±0.0237**
	$PM_{2.5}$	0.8913±0.0199	0.9154±0.0325	0.7736±0.0341	0.6938±0.0668	0.9213±0.0346	**0.9343±0.0281**
	No_x	0.5941±0.1351	0.7182±0.0973	0.7089±0.0355	0.4385±0.0832	0.7912±0.0554	**0.8140±0.0482**
25%	O_3	0.8509±0.0292	0.8891±0.0401	0.5570±0.0599	0.6106±0.0356	0.8754±0.0403	**0.9178±0.0232**
	SO_2	0.5574±0.0651	0.6368±0.0854	0.5416±0.0584	0.3994±0.0592	0.6690±0.0787	**0.7244±0.0782**
	PM_{10}	0.8574±0.0220	0.9109±0.0264	0.8555±0.0259	0.6665±0.0177	0.9167±0.0272	**0.9227±0.0251**
	$PM_{2.5}$	0.8770±0.0271	0.9121±0.0333	0.7802±0.0399	0.6917±0.0446	0.9166±0.0356	**0.9301±0.0309**
	No_x	0.6001±0.0850	0.7070±0.0989	0.6867±0.0537	0.4494±0.0954	0.7456±0.0770	**0.7958±0.0569**

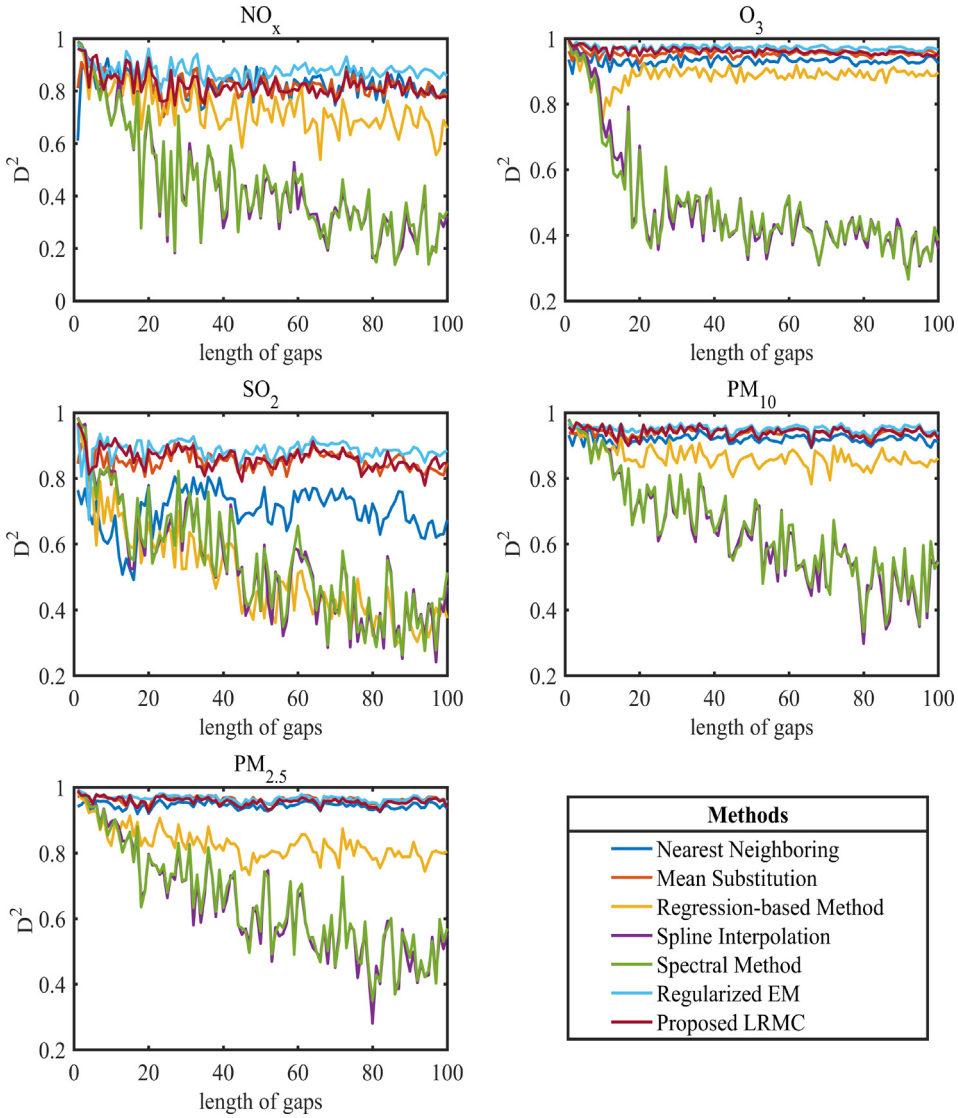

Fig. 3. D^2 comparison of imputation performance of different methods (Nearest Neighboring, Mean Substitution, Regression-based Method, Spline Interpolation, Spectral Method, Regularized EM, Proposed LRMC) under missing gaps range from 1 to 100 continuous hours.

$$\begin{aligned} &\min \quad \|A\|_* + \lambda \cdot \|E\|_1 \\ &s.\,t. \quad M^{\circ}(A + E) = M^{\circ}D \end{aligned} \tag{2}$$

Item $\|E\|_1$ is the l_1 norm of matrix E and the parameter λ controls the sparsity of noise matrix E. if E is set to the zero, it becomes to Problem 1. Alternating direction method of multipliers (ADMM) (Stephen Boyd and Chu, 2011) is applied to solve Problem 2. The major steps of the ADMM algorithm are summarized in Algorithm 1. The parameter λ entails a fine-tune process to obtain optimal imputation performance. An empirical selection is $\lambda = 1/\sqrt{\max(m, n)}$, where m, n is dimensional size of matrix D.

Algorithm 1. procedures of air-quality data imputation using low-rank matrix completion

Input: spatial matrix of air pollutant data $D \in R^{m \times n}$
set sparse weights λ, augmented lagrange coefficient ρ
initializing low-rank matrix A, missing indicator matrix M, sparse matrix E, multiplier Y
let $A = E = Y = M \Leftarrow \mathbf{0} \in \mathbf{R^{m \times n}}$
let $M(A\sim = 0) \Leftarrow 1$
while iteration not finished **do**
  singular value decomposition $[U, S, V] \Leftarrow SVD[D - E - (1/\rho)\cdot Y]$
  updating $A \Leftarrow U\cdot \mathrm{sgn}[S]^{\circ}\max[\mathrm{abs}[S] - 1/\rho, 0]\cdot V^{T}$
  updating $E \Leftarrow \mathrm{sgn}[D - A - (1/\rho)\cdot Y]^{\circ}\max[\mathrm{abs}(D - A - (1/\rho)\cdot Y - \lambda/\rho), 0]$
  updating $Y \Leftarrow Y + \rho^{\circ}M\cdot(D - A - E)$
end while
Output: $M^{\circ}D + (1 - M)^{\circ}A$

2.2. Evaluation

We use Pearson coefficient to measure the spatial correlation between different monitoring stations of various pollutants. It can be calculated as follows:

$$\rho_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3}$$

Variables $\mathbf{x}$, $\mathbf{y} \in R^{1\times n}$ are the sampling instances of two time series. n is the sampling length and $\overline{x}$, $\overline{y}$ are the average values.

In order to evaluate the imputation performance and to achieve comparable results, varieties of criteria were considered. Calculated indices of the root mean square error (*RMSE*) as follows:

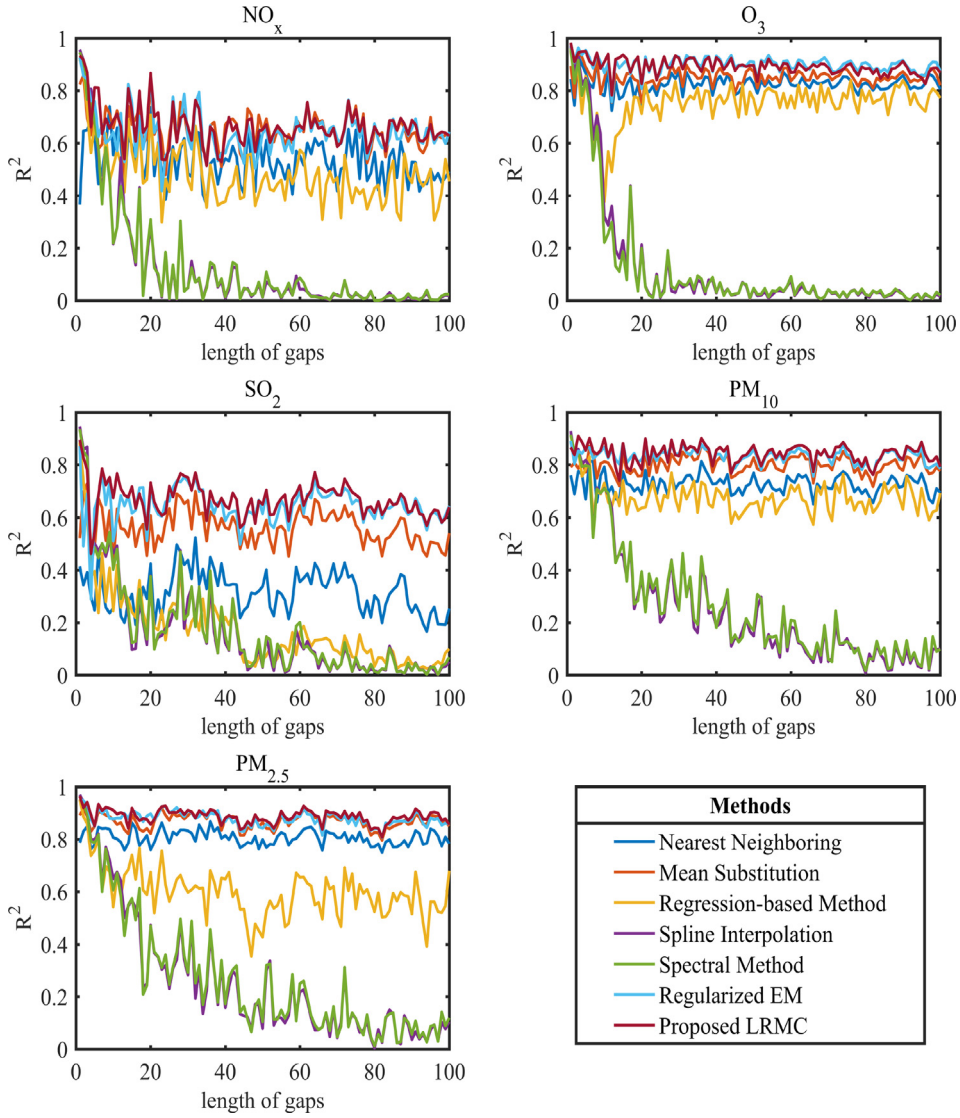

Fig. 4. R^2 comparison of imputation performance of different methods (Nearest Neighboring, Mean Substitution, Regression-based Method, Spline Interpolation, Spectral Method, Regularized EM, Proposed LRMC) under missing gaps range from 1 to 100 continuous hours.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - y_i')^2} \tag{4}$$

D^2 is agreement index, obtained as follows:

$$D^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - y_i')^2}{\sum_{i=1}^{m}(|y_i - \bar{y}| + |y_i' - \bar{y}|)^2} \tag{5}$$

and goodness of fit (R^2) is calculated as follows:

$$R^2 = \rho^2_{\mathbf{y},\mathbf{y}'} \tag{6}$$

In these equations, $\mathbf{y_i}$ denotes the observed values, $\mathbf{y_i'}$ are the imputed values, and m is the number of missing values. *RMSE* quantitatively measures average deviation between imputed values and true values which provides an evaluation from a view of single sampling point. D^2 is a standardized measure of the degree of imputation error and varies between 0 and 1. R^2 explains the similarity degree of data pattern from a view of continuous sampling series.

3. Datesets

The experimental data are collected from the AQM stations of Chang Zhou, China. They contain hourly concentration data of multiple air pollutants (NO_x, O_3, SO_2, PM_{10}, $PM_{2.5}$). These monitoring stations are located in six different sites. Geospatial distribution of six monitoring stations is shown in Fig. 1. The detailed information is showed in Table 1. We count the univariate statistical information (mean, missing-ratio, max-gap) for each air-pollutant in the year of 2016. The max missing ratio reaches 1.13% for air-pollutant O_3. The max missing gap is 34 continuous hours. The univariate statistical mean indicates that air pollutants concentration data varies slightly among stations.

4. Experiments

4.1. Data simulation

In the field of missing data analysis, the missing mechanism including missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) are often discussed from the aspect of statistical analysis (Little and Rubin, 2002). This research considers the air pollutants data are omitted in a random style because

data may be missed due to many explainable circumstances. For example, many air pollutants analyzers require one to two hours every two weeks to verify and analyze the air input flow. In addition, unexpected events that power supply failure, pump failure, electronic processor failure occur randomly that result in missingness. To better evaluate the imputation effectiveness for air pollutants dataset, the adopted simulation strategy is: given a complete spatial matrix of air pollutants data D, we remove a gap (g) of data from D obtaining $D^{\sim g}$. Then a random deletion is applied to $D^{\sim g}$ in order to assess how missing ratio influence the imputation performance. The missing gap (g) as a window slides through the whole spatial matrix, thus the simulation results can be comprehensive and convincing.

4.2. Results and analysis

In order to analyze the overall spatial correlation of pollutant data, one year length ($n = 8784$) was selected and the significance level α to compute ρ is set to 0.05. Table 2 shows the spatial correlation laws between different monitoring stations of various pollutants. It shows that the Pearson coefficients of air pollutants O_3, particle pollutants PM_{10} and $PM_{2.5}$ are greater than 0.9. That indicates that these pollutants have strong spatial correlation and are hardly affected by geographical factors and social activities (urban traffic flow, industrial production). Particle pollutants diffuse evenly in atmospheric space. Table 2 shows that the spatial correlation of pollutants decreases with the increase of the distance between monitor stations. There is not significant relation between spatial correlation and geographical distance within a certain range($\leqslant 10km$).

Given a spatial matrix of hourly pollutant concentration $D \in R^{8784\times 6}$, where $8784 = 366 \times 24$ is the number of sampling points over the whole year and the number 6 denotes the number of monitoring stations (sites A,B,C,D,E,F in Fig. 1), we obtain the estimated matrix D' under various simulated length of missing gaps, where each value was generated using the introduced imputation Algorithm 1. Therefore, an estimated matrix D' is the result of $52704 = 8784 \times 6$ runs of Algorithm 1 when the length of missing gaps is set to 1. All the simulation results are the calculated from D and D' using *RMSE*, D^2 or R^2. Fig. 2 shows the scatter plot of value pairs (t, i), where t, i denote true values and corresponding impute values under low rank matrix completion and other methods including nearest neighboring, mean substitution, regression-based method, spline interpolation, spectral method and regularized EM for comparison under the missing gap of 100 continuous hours. The imputation on pollutants O_3, PM_{10}, $PM_{2.5}$ gains better performance since scatters cluster along the read line on which the imputed values equals to the true values. It is considerably concerned with discrepancy of spatial correlation. The spatial Pearson coefficients of pollutants SO_2 and NO_x is relatively low. We can also draw from the results that if the concentration of pollutants become higher, the scatters prone to diverge from the red line. The imputation accuracy of Spectral is the worst. However, methods of LRMC, regularized EM and mean substitution generate relatively more accurate values for the corresponding true values compared to other simulated methods.

Table 3 shows the imputation results of different methods under different settings of indicators (*RMSE*, D^2 and R^2), missing ratios (5% and 10%) and air pollutants (NO_x, O_3, $PM_{2.5}$, SO_2). The length of missing gaps is 10. The best imputation results for each incomplete air pollutant data are bold. Spectral method creates the largest imputation error compared to other methods. The proposed LRMC and regularized EM algorithms are comparable and always obtain the best estimation of the missing values with indicator D^2 greater than 0.9 and R^2 greater than 0.8. Imputation results of air pollutants O_3, PM_{10}, $PM_{2.5}$ outperforms SO_2 and NO_x under all simulated methods presented in this paper. Air pollutants of SO_2 and NO_x exhibit stronger spatial variation and are more difficult to be accurately predicted. When the missing ratio rises from 5% to 10%, there is a obvious decrease in imputation precision using nearest neighboring replacement, mean substitution, regression-based method, spine interpolation and spectral method. Regularized EM and presented LRMC show a slight decrease in all indicators of *RMSE*, D^2, and R^2 if we delete more values from a given incomplete matrix of $D^{\sim g}$. Furthermore, LRMC obtains the state of the art imputation performance under the case of 10% removal from matrix of $D^{\sim g}$.

In order to access the spatial effect on LRMC based spatial, we divide the monitoring station into central station and peripheral station. Fig. 1 indicates station A is the central station, and Table 2 shows air pollutants sampled in station A have best spatial correlation than those collected in other peripheral stations B~F. Table 4 presents a summary of the imputation results of the air pollutants collected in center station A. The length of missing gaps is 10. With the missing ratio ranging from 5% to 25%, LRMC based spatial has the best performance on the indicators *RMSE*, D^2 and R^2 in most cases. The performance decrease with the increase missing ratio, but even for a missing ratio of 33%, LRMC performs comparable to regularized EM in term of indicators, *RMSE* and D^2. But in term of indicator R^2, LRMC algorithm seems always to obtain the leading imputation performance. The results of the peripheral stations shown in Table 5 indicate that the performance of LRMC in the missing data imputation of the boundary monitoring stations is worse than that of the central station. However, LRMC and EM are still the best performing algorithms, and the R^2 values of LRMC is still the highest in any case.

Figs. 3 and 4 show the imputation results using the indices of D^2 and R^2 (B) under different length of missing gaps (range from 1 to 100). From the D^2 results, temporal spline interpolation and spectral method generate lower imputation error when the length of missing gaps is less than 4. As we omit more continuous values, the D^2 of temporal interpolation and spectral method decrease significantly because the temporal structure of the air pollutant time series is unknown. Regression-based method is better than spline interpolation and spectral method but worse than other spatial imputation method including nearest neighboring replacement, mean substitution, regularized EM and LRMC which are not sensitive to the length of missing gaps. In most cases, imputation using regularized EM works better for different air pollutants compared to other methods. However, from the view of indicator R^2, LRMC algorithm seems always to obtain the leading imputation performance. It reveals that LRMC is more capable to recover the inherent trend of air pollutants time series.

5. Discussion and conclusion

In this research, we introduce a spatial imputation method which inputs the missing values using data from ambient stations accomplished by low rank matrix completion algorithm. Low rank matrix completion takes the advantage of high spatial correlation and consistency of air pollutants spatial matrix. It decomposes the spatial matrix into a low rank matrix (representing the spatial correlation) and sparse matrix (handles the probable outliers due to measurement errors), which robustly fills the unobserved values in air pollutants data sets. Thus, the pollutant space matrix have to be a low rank matrix.

Spectral method has the worst imputation performance on the air pollutants data used in this research. The imputation performance of NN, mean substitution, and regression-based method are similar for low missing ratio(5%,10%). These methods is mainly related to length of gaps and low missing ratio. The effect of these factors on imputation performance of these methods has already been studied in literature (Junninen et al., 2004; Moshenberg et al., 2015). However, LRMC performs comparable to regularized EM in term of indicators, *RMSE* and D^2, the accuracy and agreement of those are very good. In term of indicator R2, LRMC algorithm seems always to obtain the leading imputation performance. LRMC optimally recovers the inherent data structure (trend) compared to other spatial imputation methods with respect to goodness of fit (R^2). LRMC can be a better choice to deal with

long missing gaps with a certain ratio of missing values.

For air pollutants O_3, PM_{10}, $PM_{2.5}$, spatial correlation of those is stronger than air pollutants SO_2 and NO_x. Compared to O_3, the chemical lifetime of NO is shorter and relative spatial inhomogeneity(Shahbazi et al., 2018). Furthermore, correlation is affected by spatial distance (Tabel 2) of the monitoring stations. The imputation performance of LRMC is affected by spatial correlation. For the stronger spatial correlated air pollutant, the imputation performance is better. Thus, central station has better performance than the boundary stations.

In our future work, we will investigate how we can further evaluate the imputation performance with more advanced imputation evaluation framework, for example, the work of (Chapman et al., 2018) proposed a framework to evaluate imputation performance for every imputation method.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part National key research and development program 2018AAA0100800, by the Key Research and Development Program of Jiangsu under grants BK20192004, BE2018004-04, BE2017071, and BE2017647, by the Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University under grant 2019005; and by the State Key Laboratory of Integrated Management of Pest Insects and Rodents (Grant No.IPM1914).

References

Baldauf, R.W., Lane, D.D., Marote, G.A., 2001. Ambient air quality monitoring network design for assessing human health impacts from exposures to airborne contaminants. Environ. Monit. Assess. 66 (1), 63. https://doi.org/10.1023/a:1026428214799.

Box, G.E.P., 2008. Time series analysis: Forecasting and control. 4th ed., J. Market. Res. 14 (2). https://doi.org/10.2307/3150485.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. Lancet 360 (9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Mathe. 9 (6), 717. https://doi.org/10.1007/s10208-009-9045-5.

Chandrappa, R., Kulshrestha, U.C., 2016. Major Issues of Air Pollution. Springer International Publishinghttps://doi.org/10.1007/978-3-319-21596-9_1.

Chapman, A., Pang, W., Coghill, G., 2018. CLEMI-imputation evaluation, 000373–000378.

Gómez-Carracedo, M.P., Andrade, J.M., López-Mahía, P., Muniategui, S., Prada, D., 2014. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemomet. Intell. Lab. Syst. 134 (8), 23–33. https://doi.org/10.1016/j.chemolab.2014.02.007.

Guan, Q.Y., 2016. Judgment and handling of abnormal data during ambient air automatic monitoring data audit. Environ. Monit. Forewarning.

H Jr., W.F., Beebe, R.C., 1985. Quality assurance in air pollution measurements. Air Repair 29 (7), 699–700.

Junger, W.L., Leon, A.P.D., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ. 38 (18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026.

Listed, N., 2014. 7 million deaths annually linked to air pollution. Cent. Eur. J. Public Health 22 (1), 53–59.

Little, R.J., Rubin, D.B., 2002. Statistical analysis with missing data. Technometrics 45 (4), 364–365. https://doi.org/10.1002/9781119013563.ch10.

Mansour, S., Farshad, F., Kimiya, G., Mahboubeh, P., Hassan, A., Katayoun, R., Mohammad Sadegh, H., Iman, N., Akbar, F., Kazem, N., 2014. A framework for exploration and cleaning of environmental data–tehran air quality data experience. Arch. Iranian Med. 17 (12), 821–829 https://doi.org/10.0141712/AIM.008.

Moshenberg, S., Lerner, U., Fishbain, B., 2015. Spectral methods for imputation of missing air quality data. Environ. Syst. Res. 4 (1), 1–13. https://doi.org/10.1186/s40068-015-0052-z.

Noor, N.M., 2015. Filling the missing data of air pollutant concentration using single imputation methods. Appl. Mech. Mater. 754–755, 923–932. https://doi.org/10.4028/www.scientific.net/AMM.754-755.923.

Plaia, A., Bondì, A., 2006. Single imputation method of missing values in environmental pollution data sets. Atmos. Environ. 40 (38), 7316–7330. https://doi.org/10.1016/j.atmosenv.2006.06.040.

Pope III, C.A., 2000. Epidemiological basis for particulate air pollution health standards. Aerosol Sci. Technol. 32 (1), 4–14. https://doi.org/10.1080/027868200303885.

Schneider, T., 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J. Climate 14 (5), 853–871.

Seinfeld, J.H., Pandis, 2006. Atmospheric chemistry and physics: From air pollution to climate change.

Shafer, J.L, 1997. Analysis of Incomplete Multivariate Data.

Shahbazi, H., Karimi, S., Hosseini, V., Yazgi, D., Torbatian, S., 2018. A novel regression imputation framework for tehran air pollution monitoring network using outputs from wrf and camx models. Atmos. Environ. https://doi.org/10.1016/j.atmosenv.2018.05.055. S1352231018303650.

Shang, F., Liu, Y., Cheng, J., Cheng, H., 2014. Robust principal component analysis with missing data, https://doi.org/10.1145/2661829.2662083.

Shukri, Y.A., Noraziana, M.N., Al, A.M.M., 2008. Estimation of missing values in air pollution data using single imputation techniques. Scienceasia 34 (3), 341–345. https://doi.org/10.2306/scienceasia1513-1874.2008.34.341.

Stephen Boyd, N.P., Chu, E., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers 3(1), 1–22. https://doi.org/10.1561/2200000016.

von Lehmden, D., DeWees, W., Nelson, C., 1979. Quality assurance handbook for air pollution measurement systems. Volume iii. Stationary source specific methods. Tech. Rep., Environmental Protection Agency, Research Triangle Park, NC (USA).

Wang, L., Yang, Z., Kai, W., Bo, Z., Qiang, Z., Wei, W., 2014. Application of weather research and forecasting model with chemistry (wrf/chem) over northern china: Sensitivity study, comparative evaluation, and policy implications. Atmos. Environ. 124, 337–350. https://doi.org/10.1016/j.atmosenv.2014.12.052.

Wright, J., Ganesh, A., Rao, S., Ma, Y., 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices, 58(3), 289–298 vol. 1. https://doi.org/10.1109/NNSP.2000.889420.

Yaroslavsky, L.P., Gil, S., Salomon, B.G., Ideses, I.A., Barak, F., 2009. Nonuniform sampling, image recovery from sparse data and the discrete sampling theorem. J. Opt. Soc. Am. A Opt. Image Sci. Version 26 (3), 566–575. https://doi.org/10.1364/josaa.26.000566.